Modeling Frequency and Severity of Claims with the Generalized Cluster-Weighted Model

Nikola Počuča, Tatjana Miljkovic, Petar Jevtić, Paul McNicholas

August 20, 2018

Abstract

In this paper, we propose a generalized cluster-weighted model (GCWM) that allows for modeling non-Gaussian distribution of the continuous covariates and a new zero-inflated GCWM (ZI-GCWM) for modeling insurance claims data with excess zeros. We describe two expectation-optimization (EM) algorithms for parameter estimation in GCWM and ZI-GCWM. A simulation study showed that both cluster models perform well for different settings in contrast to the existing mixture-based approaches. A real data set based on French automobile policies is used to illustrate the application of the proposed models.

KEY WORDS: GCWM, CWM, ZI-GCWM, clustering, automobile claims.
JEL CLASSIFICATION: C02, C40, C60.

1 Introduction

A significant number of clustering methods have been proposed for sub-grouping the data in the area of computer science, biology, social science, statistics, marketing, etc. Ingrassia and Minotti (2015) proposed a cluster-weighted models (CWMs) as a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of mixed-type covariates with the assumption that continues covariates come from Gaussian distribution. The CWM models with Gaussian assumptions have been proposed by Gershenfeld (1997), Gershenfeld and Metois (1999), and Gershenfeld (1999) in a context of media technology. Some extensions of this class of models have been considered by Punzo and Ingrassia (2014), Ingrassia and Punzo (2014), and Ingrassia and Vittadini (2014). These clustering methods luck some capabilities in order to be able to accommodate modeling insurance data (e.g. high excess zeros for claim count, heavy-tail loss distribution, deductible, or limits).

Sub-grouping of insurance policies based on risk classification is a standard practice in insurance. The heterogenous nature of insurance data allows for explorations of many different techniques for sub-grouping risk. As a result, there is a growing number of papers in the area of mixture modeling of univariate and multivariate insurance data to account for heterogeneity of risk. Lee and Lin (2010), Verbelen et al. (2015), and Miljkovic and Grün (2016) proposed mixture models for univariate loss data. The idea of mixture modeling of univariate insurance data has been extended to a multivariate classification. A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking was studied by Bermúdez and Karlis (2012). A Poisson mixture model for count data was considered by Brown and Buckley (2015) with application in managing Group Life insurance portfolio. Miljkovic and Fernndez (2018) reviewed two complementary mixture-based clustering approaches (cluster-weighted model and mixture-based clustering for an ordered stereotype model) for modeling unobserved heterogeneity in an automobile insurance portfolio, depending on the data structure under consideration. Mixture models with the applications is financial mathematics have been explored by **?**, **?**, and many others.

In this paper paper, we extend the CWM family proposed by Ingrassia and Minotti (2015) to allow for modeling of non-Gaussian continues covariates and a zero-inflated Poisson (ZIP) claims data with excess zeros which are commonly seen in the insurance applications. We define our proposed model as the generalized cluster-weighted model as GCWM and a new zero-inflated GCWM as ZI-GCWM. Two partitioning methods are considered with two separate EM algorithms. The first EM algorithm is for generating the GCWM models, while the second EM is for optimizing the ZI-GCWM. We show that the Bernoulli and Poisson GCWM accurately estimate the initialization of the EM algorithm for the ZI-GCWM model. These models utilize individual policy and claims data and should be useful in the areas of ratemaking and risk management.

This paper is organized as follows. Section 2 presents the proposed model for mixture of GLMs. Section 3 applies the proposed model on a real data of French automobile claims. An extensive simulation study is discussed in Section 4. Conclusion is provided in Section 5.

2 Proposed Model

2.1 Background

Let $(\boldsymbol{X}', Y)'$ be the pair of a vector of covariates $\boldsymbol{X}$ and a response variable Y. Assume this set is defined on some space Ω that takes values in appropriate Euclidian subspace. Now, assume that there exists G partitions of

Ω, denoted as $\Omega_1, \ldots, \Omega_G$. Gershenfeld (1997) characterized the Cluster weighted models as a finite mixture of GLMs hence, the joint distribution $f(\boldsymbol{x}, y)$ of $(\boldsymbol{X}^{'}, Y)^{'}$ is expressed as follows

$$f(\boldsymbol{x}, y) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \Omega_j) p(\boldsymbol{x}; \Omega_j). \tag{2.1}$$

The pair $q(y|\boldsymbol{x}; \Omega_j)$ and $p(\boldsymbol{x}; \Omega_j)$ are conditional and marginal distributions of $(\boldsymbol{X}^{'}, Y)^{'}$ respectively, while τ_j represents the weight of the jth component such that $\sum_{j=1}^{G} \tau_j = 1$, $\tau_j > 0$. Ingrassia and Minotti (2015) proposed a flexible family of mixture models for fitting the joint distribution of a random vector $(\boldsymbol{X}^{'}, Y)^{'}$ by splitting the covariates into continues and discrete as $\boldsymbol{X} = (\boldsymbol{V}^{'}, \boldsymbol{W}^{'})^{'}$. This assumption of independence between continues and discrete covariates allows us to multiply their corresponding marginal distributions. Thus, for this setting the model in (2.1) is reformulated as follows

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) p(\boldsymbol{x}; \boldsymbol{\theta_j}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) p(\boldsymbol{v}; \boldsymbol{\theta^{\star}_j}) p(\boldsymbol{w}; \boldsymbol{\theta^{\star\star}_j}) \tag{2.2}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the vectors of continues and discrete covariates respectively, the $q(y|\boldsymbol{x}; \boldsymbol{\vartheta_j})$ is a conditional density of $Y|\boldsymbol{x}$, with parameter vector $\boldsymbol{\vartheta_j}$, the $p(\boldsymbol{v}; \boldsymbol{\theta^{\star}_j})$ is the marginal distribution of $\boldsymbol{v}$ with parameter vector $\boldsymbol{\theta^{\star}_j}$. the $p(\boldsymbol{w}; \boldsymbol{\theta^{\star\star}_j})$ is the marginal distribution of $\boldsymbol{w}$ with parameter vector $\boldsymbol{\theta^{\star\star}_j}$. Finally $\boldsymbol{\Phi} := (\boldsymbol{\theta}^{\star}, \boldsymbol{\theta}^{\star\star}, \boldsymbol{\tau}, \boldsymbol{\vartheta})$ includes all model parameters. In addition, the conditional distribution $q(y|\boldsymbol{x}; \boldsymbol{\vartheta_j})$ is assumed to belong to an exponential family of distributions and as such can be modeled in the framework of GLMs. Here, the marginal distribution of continues covariates is assumed to be of Gaussian type. Unfortunately, this last assumption is too strong for use in insurance related applications specifically in rate-making. To relax it, we develop the Generalized cluster-weighted mdoel (GCWM) that allows for non-Gaussian covariates as discussed in the next section.

2.2 Generalized cluster-weighted model (GCWM)

We proceed to extend (2.2) by splitting the continues covariates further as $\boldsymbol{V} := (\boldsymbol{U}^{'}, \boldsymbol{T}^{'})^{'}$, where $\boldsymbol{U}$ is a set of non-Gaussian covariates, and $\boldsymbol{T}$ a set of Gaussian covariates. Thus (2.2) is now recovered as

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) p(\boldsymbol{t}; \boldsymbol{\theta^{\star}_j}) p(\boldsymbol{w}; \boldsymbol{\theta^{\star\star}_j}) p(\boldsymbol{u}; \boldsymbol{\theta^{\star\star\star}_j}) \tag{2.3}$$

which we refer to as generalized cluster-weighted model (GCWM). Here $p(\boldsymbol{t}; \boldsymbol{\theta^{\star}_j})$ denotes the marginal density of Gaussian covariates, with parameter vector $\boldsymbol{\theta}^{\star}$, and $p(\boldsymbol{u}; \boldsymbol{\theta^{\star\star\star}_j})$ as the marginal density of the non-Gaussian covariates with parameter vector $\boldsymbol{\theta^{\star\star\star}_j}$.

As it is relevant to the actuarial application in this paper, we focus on the multivariate log-normal distribution for non-Gaussian covariates. This however does not reduce the generality of our approach. Now, with our log-normal assumption for $p(\boldsymbol{u};\boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star})$, we have that $\boldsymbol{u}$ is defined on $\mathbb{R}_+$ with parameter vector $\boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star}$, and having probability density function as

$$p\left(\boldsymbol{u};\boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star} := (\boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star},\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\star\star\star})\right) = \frac{1}{(\prod_{i=1}^{N} u_i)|\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\star\star\star}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star})^{\prime}\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\star\star\star -1}(\ln \boldsymbol{u} - \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star})\right]. \tag{2.4}$$

The derivation of the equation above can be found in the Appendix 7.1.

2.3 Zero-inflated Poisson

For the zero-inflated Poisson model (ZIP) see Lambert (1992) we can split the conditional density $p(y|\boldsymbol{x},\boldsymbol{\vartheta}_{\boldsymbol{j}})$ into zero and non-zero densities. The response y variable when $y=0$ are distributed with density $q(0|\boldsymbol{x},\boldsymbol{\vartheta}_{\boldsymbol{j}})$. The response y values when $y>0$ are distributed with density $q(y>0|\boldsymbol{x},\boldsymbol{\vartheta}_{\boldsymbol{j}})$. Given the conditional density now defined for the ZIP model, (2.3) can be re-written as follows

$$f(\boldsymbol{x},y;\Phi) = \sum_{j=1}^{G} \tau_j \left[q(0|\boldsymbol{x};\boldsymbol{\vartheta}_{\boldsymbol{j}}) + q(y>0|\boldsymbol{x};\boldsymbol{\vartheta}_{\boldsymbol{j}}))\right] p(\boldsymbol{t};\boldsymbol{\theta}_{\boldsymbol{j}}^{\star})p(\boldsymbol{w};\boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star})p(\boldsymbol{u};\boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star}). \tag{2.5}$$

Let $\tilde{\boldsymbol{x}} := [\boldsymbol{1},\boldsymbol{x}]$, where $\tilde{\boldsymbol{x}}$ is a matrix of covariates with the addition of a placeholder for the intercept in the GLM. We denote the Poisson conditional density as $q^P(y|\boldsymbol{x};\boldsymbol{\beta}_{\boldsymbol{j}})$, where $y \in \{0,1,\dots\}$, and $\boldsymbol{\beta}_{\boldsymbol{j}}$ is the row coefficient vector. Here, the link function will be modelled with log-link for the GLM:

$$\lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\boldsymbol{j}}^{\prime}}, \qquad\qquad q^P(y|x;\lambda_j) = e^{-\lambda_j}\frac{{\lambda_j}^y}{y!}.$$

Next, we introduce a Bernoulli process for the conditional density. We denote the density as $q^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_{\boldsymbol{j}})$, where $\mathbb{1}(\cdot)$ is the indicator function and $\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}$ is the coefficient vector. Here, the GLM will be modeled with the associated logit link function

$$\psi_j = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}^{\prime}}}{1+e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}^{\prime}}}, \qquad\qquad q^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}) = \begin{cases} \psi_j, & \mathbb{1}(y)=0 \\ 1-\psi_j, & \mathbb{1}(y)=1 \end{cases}$$

Now, given a combination of two preceding models, we introduce the ZIP process in which zero counts come from two random variables. One comes from Bernoulli random variable which generates structural zeros, and the other comes from the Poisson random variable. The coefficients $\boldsymbol{\vartheta}_{\boldsymbol{j}} := \{\boldsymbol{\beta}_{\boldsymbol{j}},\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}\}$ correspond to the two above introduced conditional densities where the coefficients are estimated using a generalized linear model as in Lambert (1992). The components of ZIP conditional density $q(y|\boldsymbol{x};\boldsymbol{\vartheta_j})$ are

$$q(0|x;\boldsymbol{\vartheta_j}) = \psi_j + (1-\psi_j)e^{-\lambda_j} \qquad \text{and} \qquad q(y>0|x;\boldsymbol{\vartheta_j}) = (1-\psi_j)e^{-\lambda_j}\frac{(\lambda_j)^y}{y!}.$$

Also, the link functions to consider are log-link for the Poisson model and logit link for the Bernoulli

$$\psi_j = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}{}'}}{1+e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}{}'}} \qquad \text{and} \qquad \lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}'_{\boldsymbol{j}}}.$$

Let paramater ψ_j denote the probability that the zero comes from the Bernoulli distribution of jth component, and the parameter λ_j characterizes the jth Poisson distribution. This allows for a more nuanced approach to handling the inflation of zeros similarly as in Bermúdez and Karlis (2012).

3 Introducing Bernoulli-Poisson Partitioning

The single component ZIP model assumes that the inflated zeros emanate from both a Bernoulli and Poisson random variables while the non-zeros are assumed to come exclusively from the Poisson random variable. However, recent research extends the single component ZIP models to mixture models for heterogeneous count data with excess zeros (see (Bermúdez and Karlis, 2012)). In mixtures of ZIPs, zeros are assumed to come from multiple different Binomial and Poisson random variables. Difficulties are apparent during the maximization step of the EM when means of covariates are very close together (see Lim et al. (2014)). However, misclassification error can be reduced using parsimonious models for the independent variables as in McNicholas et al. (2010).

In this work, we propose a new method to rectify this problem and partition the dataset using Bernoulli and Poisson GCWMs. Furthermore, we construct a new zero inflated GCWM (ZI-GCWM) using the previously generated Bernoulli and Poisson GCWMs. We show that the Bernoulli and Poisson GCWM accurately estimate the initialization of the EM algorithm for the zero inflated GCWM model. The work of Lambert (1992) specifies that the MLE estimates for coefficients provide an excellent guess allowing EM to converge quickly for ZIPs. The partitioning method consists of two separate EM algorithms. The first EM algorithm is for generating the GCWM models, while the second EM is for optimizing the ZI-GCWM. Recall $(\boldsymbol{X'}, Y)'$ to be a vector defined on some sample space Ω. As discussed, this sample space is partitioned into G non-overlapping sets such that their union constitutes this sample space ie. $\Omega = \bigcup_{i=1}^{G} \Omega_i$. However, contingent on a model choice each particular set Ω_i may take a different shape. Specifically, if we introduce the Bernoulli model in a generalized form for conditional density (see Ingrassia and Minotti (2015) for specific cases), we have the sample space

Ω^B and joint pdf f^B to be

$$\Omega^B = \bigcup_{l=1}^{G} \Omega_l^B \qquad \text{and} \qquad f^B(\boldsymbol{x}, y; \Phi) = \sum_{l=1}^{G} \tau_l q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_{\boldsymbol{l}}) p(\boldsymbol{t}; \boldsymbol{\theta}_{\boldsymbol{l}}^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_{\boldsymbol{l}}^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_{\boldsymbol{l}}^{\star\star\star}).$$

Similarly if we introduce a Poisson model in a generalized form the sample space Ω^P and joint pdf f^P become

$$\Omega^P = \bigcup_{j=1}^{M} \Omega_j^P \qquad \text{and} \qquad f^P(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{M} \tau_j q^P(y|\boldsymbol{x}; \boldsymbol{\beta}_{\boldsymbol{j}}) p(\boldsymbol{t}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star}).$$

Where this sample space is partitioned up to M non-overlapping sets. Now, construct a new partitioning of a sample space Ω such that

$$\Omega = \Omega^Z = \bigcup_{l\in\{1\ldots G\}, j\in\{1\ldots M\}} \Omega_{l,j}^Z := \bigcup_{l\in\{1\ldots G\}, j\in\{1\ldots M\}} \Omega_l^B \cap \Omega_j^P =: \bigcup_{k\in\{1\ldots K\}} \Omega_k^Z$$

, where K can range up to $M \times G$ unique partitions. Therefore the new conditional density is now result of a model in which each component is captured by the conditional probability density function that is of mixture of particular Bernoulli and particular Poisson

$$q_k^Z(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k) := q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k) + (1 - q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k)) q^P(y|\boldsymbol{x}; \boldsymbol{\beta}_k). \tag{3.1}$$

$$= q(0|\boldsymbol{x}; \boldsymbol{\vartheta}_{\boldsymbol{k}}) + q(y > 0|\boldsymbol{x}; \boldsymbol{\vartheta}_{\boldsymbol{k}}) \tag{3.2}$$

The expectation-maximization (EM) algorithm (see Dempster et al. (1977)) is then used to estimate this new mixture of up to $M \times G$ specific GCWMs. The initialization parameters for the second EM algorithm are provided by Bernoulli and Poisson GCWMs from (3.1) giving parameter pairs (ψ_k, λ_k) where $k \in \{1 \ldots K\}$. The second EM procedure then optimizes (3.2). The ZI-GCWM is then compared against the standard Poisson GCWM using a liklihood ratio test which is commented in section 3.6.

3.1 The EM Algorithm for Parameter Estimation

In most finite mixture problems, the standard method for estimating parameters of mixture models is based on the EM algorithm and further discussed by McLachlan and Peel (2000). The Bernoulli-Poisson partitioning method is split into two EM algorithm steps. The first EM partitions the sample space, while the second EM optimizes the zero inflated portion.

3.2 EM - Partitioning

The EM algorithm is based on the local maximum likelihood estimation. The initial values of the parameter estimates can be generated from a variety of strategies outlined in Biernacki et al. (2000), then the algorithm proceeds by alternation of the E-step and M-step to update parameter estimates. To find an optimal number of components, maximum likelihood estimation is obtained over a range of G, and the best model is selected based on the Bayesian information criterion (BIC).

The convergence criterion of the EM algorithm is based on the Aitken acceleration. It is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm. when the relative increase in the log-likelihood function is no bigger than a small pre-specified tolerance value or the number of iterations reach a limit. In this subsection, we explain the parameter estimation in line with the GCWM methodology proposed by Ingrassia and Minotti (2015). The proposed GCWM a model is based on the assumption that $q(y|\boldsymbol{x}, \vartheta_j)$ belongs to the exponential family of distributions that are strictly related to GLMs. The link function relates the expected value $g(\mu_j) = \beta_{0j} + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p$. We are interested in estimation of the vector $\boldsymbol{\beta}_j$, thus the distribution of $y|\boldsymbol{x}$ is denoted by $q(y|\boldsymbol{x}; \boldsymbol{\beta_j}, \lambda_j)$, where λ_j denotes an additional parameter to account for when a distribution belong to a two-parameter exponential family.

The marginal distribution $p(\boldsymbol{x}; \boldsymbol{\theta}_j)$ has the following components: $p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star})$, $p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})$, and $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$. The first marginal density $p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star})$ is modeled as a Gaussian with mean $\boldsymbol{\mu_j}$ and covariance matrix $\boldsymbol{\Sigma}_j$ as $p(\boldsymbol{t}; \boldsymbol{\mu_j}, \boldsymbol{\Sigma}_j)$. The marginal density $p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})$ assume that each finite discrete covariate W is represented as a vector $\boldsymbol{w}^r = (w^{r1}, \ldots, \boldsymbol{w}^{rc_r})'$ where $w^{rs} = 1$ if $w_r = s$, such that $s \in \{1, \ldots, c_r\}$, and $w^{rs} = 0$ otherwise.

$$p(\boldsymbol{w}, \boldsymbol{\gamma_j}) = \prod_{r=1}^{d} \prod_{s=1}^{c_r} (\gamma_{jrs})^{w^{rs}} \tag{3.3}$$

for $j = 1, \ldots, G$, where $\boldsymbol{\gamma_j} = (\gamma_{j1}^{'}, \ldots, \gamma_{jd}^{'})'$, $\boldsymbol{\gamma}_{jr} = (\gamma_{jr1}^{'}, \ldots, \gamma_{jrc_d}^{'})'$, $\gamma_{jrs} > 0$, and $\sum_{s=1}^{c_r} \gamma_{jrs}$, $r = 1, \ldots, q$. The density $p(\boldsymbol{w}, \boldsymbol{\gamma_j})$ represents the product of d conditionally independent multinomial distributions with parameters $\boldsymbol{\gamma}_{jr}$, $r = 1, \ldots, d$. The last marginal density $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$ will be modelled as lognormal with mean vector $\boldsymbol{\mu}_j^{\star\star\star}$, and covariance matrix $\boldsymbol{\Sigma}_j^{\star\star\star}$.

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of n independent observations drawn from model in (2.3). For this sample, the complete data likelihood function, $L_c(\boldsymbol{\Phi})$, is given by

$$L_c(\boldsymbol{\Phi}) = \prod_{i=1}^{n} \prod_{j=1}^{G} \left[\tau_j q(y_i|x_i, \boldsymbol{\beta}_j, \lambda_j) p(t_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) p(w_i, \gamma_j) p(u_i, \boldsymbol{\mu_j^{\star\star\star}}, \boldsymbol{\Sigma_j^{\star\star\star}})\right]^{z_{ij}}, \tag{3.4}$$

where z_{ig} is the latent indicator variable with value of $z_{ig} = 1$ indicating that observation $(\boldsymbol{x_i}, y_i)$, originated from the jth mixture component and $z_{ij} = 0$ otherwise.

By taking the logarithm of (3.4), the complete-data log-likelihood function $\ell_c(\mathbf{\Phi})$ is written by

$$\ell_c(\mathbf{\Phi}) = \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\big[\log(\tau_j) + \log q(y_i|x_i, \boldsymbol{\beta}_j, \lambda_j) + \log p(t_i, \boldsymbol{\mu}_j, \mathbf{\Sigma}_j) + \log p(w_i, \gamma_j) + \log p(u_i, \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star}, \mathbf{\Sigma}_{\boldsymbol{j}}^{\star\star\star})\big].$$

3.2.1 E-Step - Partitioning

The E-step does not depend on the form of density, and the latent data only relate to $\boldsymbol{z}$. The posterior probability that $(\boldsymbol{x_i}, y_i)$ comes from the jth mixture component is calculated at the sth iteration of the EM algorithm as

$$\begin{aligned}\pi_{ij}{}^{(s)} &= E[z_{ij}|(\boldsymbol{x_i}, y_i), \mathbf{\Phi}^{(s)}] \\ &= \frac{\tau_j{}^{(s)} q(y_i|x_i, \boldsymbol{\beta}_j^{(s)}, \lambda_j^{(s)}) p(t_i, \boldsymbol{\mu}_j^{(s)}, \mathbf{\Sigma}_j^{(s)}) p(w_i, \gamma_j^{(s)}) p(u_i, \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star(s)}, \mathbf{\Sigma}_{\boldsymbol{j}}^{\star\star\star(s)})}{f(\boldsymbol{x_i}, y_i; \Phi^{(s)})}.\end{aligned}$$

3.2.2 M-Step - Partitioning

It follows that at the $(s+1)$th iteration, the conditional expectation of (3.2) on the observed data and the estimates from the sth iteration results in

$$\begin{aligned}Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)}) &= \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)}\big[\log(\tau_j) + \log q(y_i|x_i, \boldsymbol{\beta}_j, \lambda_j) + \log p(t_i, \boldsymbol{\mu}_j, \mathbf{\Sigma}_j) + \log p(w_i, \gamma_j) + \log p(u_i, \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star(s)}, \mathbf{\Sigma}_{\boldsymbol{j}}^{\star\star\star(s)})\big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)} \log(\tau_j) + \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)} \log q(y_i|x_i, \boldsymbol{\beta}_j^{(s)}, \lambda_j^{(s)}) + \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)} \log p(t_i, \boldsymbol{\mu}_j^{(s)}, \mathbf{\Sigma}_j^{(s)}) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)} \log p(w_i, \gamma_j^{(s)}) + \sum_{i=1}^{n}\sum_{j=1}^{G} \pi_{ij}^{(s)} \log p(u_i, \boldsymbol{\mu}_{\boldsymbol{j}}^{\star\star\star(s)}, \mathbf{\Sigma}_{\boldsymbol{j}}^{\star\star\star(s)}). \quad (3.5)\end{aligned}$$

The M-step requires maximization of the Q-function with respect to $\mathbf{\Phi}$ which can be done separately for each term on the right hand side in (3.5). As a result, the parameter updates $\hat{\tau}_j$, $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\sigma}}_j$, and $\hat{\boldsymbol{\gamma}}_j$ on the $(s+1)$th iteration are:

$$\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} \boldsymbol{t}_i, \qquad \hat{\boldsymbol{\gamma}}_{jr}^{(s+1)} = \frac{\sum_{i=1}^{n} \pi_{ij}^{(s)} \omega_i^{rs}}{\sum_{i=1}^{n} \pi_{ij}^{(s)}},$$

$$\hat{\boldsymbol{\sigma}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} (\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})',$$

The log-normal distribution is relevant for modelling actuarial data. Parameter estimates for the log-normal distribution follow similar suit.

$$\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} \ln \boldsymbol{u}_i,$$

$$\hat{\boldsymbol{\sigma}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^n \pi_{ij}^{(s)}} \sum_{i=1}^n \pi_{ij}^{(s)} (\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})(\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})'.$$

The estimates of $\boldsymbol{\beta}$ are computed by maximizing each of the G terms

$$\sum_{i=1}^n \pi_{ij}^{(s)} \log q(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_j, \lambda_j). \tag{3.6}$$

Maximization of (3.6) is performed by numerical optimization in R software in a similar framework the mixture of generalized linear models are implemented. For additional details about this implementation the reader is refer to Wedel and De Sabro (1995) and Wedel (2002). For insurance applications, current TCWM model can be used for modeling frequency of claims assuming that $\boldsymbol{Y}$ belongs to Poisson or Bernoulli distributions. When modelling severity of claims, $\boldsymbol{Y}$ can be assumed accommodate Gamma or Lognormal distributions. All of these applications are based on CWM as the underlying approach. For additional information, the reader is referred to the manual of the `flexCWM` package manual for $\mathbf{R}$ users written by Ingrassia and Minotti (2015).

3.3 EM - Zero-inflated

The optimization of the zero-inflated model (3.2) uses the EM algorithm for maximizing the incomplete-data log-liklihood iteratively (Lambert, 1992). The log-likilhood for ψ_k and λ_k is expressed as

$$l(\psi_k, \lambda_k; y, \boldsymbol{x}) = \sum_{y_i=0} \log \left[e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{\beta}}_{\boldsymbol{k}}'} + \exp\left(-e^{-\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \boldsymbol{\beta}_{\boldsymbol{k}}'}\right)\right] + \sum_{y_i>0} \left(y_i \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_{\boldsymbol{k}}' + e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_{\boldsymbol{k}}'}\right) - \sum_{i=1}^n \log\left(1 + e^{\tilde{\boldsymbol{x}}_i \bar{\beta}_k{}'}\right) - \sum_{y_i>0} \log(y_i!)$$

Where y_i, and $\tilde{\boldsymbol{x}}_i$ referes to the ith row of the response variable y and covariate matrix $\tilde{\boldsymbol{x}}$. Due to the first term, the log-liklihood is complicated to maximize, Lambert (1992) provides a meaningful solution. Suppose that we could observe $Z_{ik} = 1$ when y_i is generated from the Bernoulli random variable of partition k, and $Z_{ik} = 0$ when y_i is generated from the Poisson random variable. Then the complete-data log-liklihood would be written as

$$\begin{aligned} l_c(\psi_k, \lambda_k; y, \boldsymbol{z_k}) =& \sum_{i=1}^n \left(z_{ik} \tilde{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{\beta}}_{\boldsymbol{k}}{}' - \log(1 + e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{\beta}}_{\boldsymbol{k}}{}'}\right) + \sum_{i=1}^n (1 - z_{ik})(y_i \tilde{\boldsymbol{x}}_{\boldsymbol{i}} \boldsymbol{\beta}_{\boldsymbol{k}}' - e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \boldsymbol{\beta}_{\boldsymbol{k}}'}) + \sum_{i=1}^n (1 - z_{ik}) \log(y_i!) \\ =& \quad l_c(\psi_k; \boldsymbol{y}, \boldsymbol{z_k}) + l_c(\lambda_k; \boldsymbol{y}, \boldsymbol{z_k}) + \sum_{i=1}^n (1 - z_{ik}) \log(y_i!) \end{aligned} \tag{3.7}$$

where z_{ik} is a realization of Z_{ik}. (3.7) is easier to maximize since $l_c(\psi_k; \boldsymbol{y}, \boldsymbol{z_k})$ and $l_c(\lambda_k; \boldsymbol{y}, \boldsymbol{z_k})$ can be maximized seperately for parameters λ_k and ψ_k. With the EM algorithm, the incomplete-data log-liklihood can be maximized iteratively between estimating Z_{ik} with its expectation under current parameters λ_k and ψ_k

(E-Step) and then maximizing the complete data-logliklihood (M-Step).

3.3.1 E-step - Zero-inflated

Using current estimates $\psi_k^{(s)}$ and $\lambda_k^{(s)}$ from the partition Ω_{Z_k}, we calculate the expected value of Z_{ik} by its posterior mean $z_{ik}^{(s)}$ for each cluster k, at iteration s

$$z_{ik}^{(s)} = \begin{cases} \Big[1 + \exp\big(-\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\bar{\boldsymbol{\beta}}_{\boldsymbol{k}}'^{(\boldsymbol{s})} - e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta}_{\boldsymbol{k}}'^{(\boldsymbol{s})}}\big)\Big]^{-1}, & y_i = 0 \\ 0 \quad , & y_i > 0. \end{cases}$$

3.3.2 M-Step - Zero-inflated

The M-Step can be split into the maximization of two complete data log-likelihoods and the $\boldsymbol{z_i}$ calculated from the previous iteration (s) as:

$$l_c(\psi_k; \boldsymbol{y}, \boldsymbol{z}_{\boldsymbol{k}}^{(\boldsymbol{s})}) = \sum_{i=1}^{n} \Big(z_{ik}^{(s)}\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\bar{\boldsymbol{\beta}}_{\boldsymbol{k}}'^{(\boldsymbol{s})} - \log(1 + e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\bar{\boldsymbol{\beta}}_{\boldsymbol{k}}'^{(s)}}\Big) \tag{3.8}$$

$$l_c(\lambda_k; \boldsymbol{y}, \boldsymbol{z}_{\boldsymbol{k}}^{(\boldsymbol{s})}) = \sum_{i=1}^{n}(1 - z_{ik}^{(s)})(y_i\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta}_{\boldsymbol{k}}'^{(\boldsymbol{s})} - e^{\tilde{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta}_{\boldsymbol{k}}'^{(s)}}) \tag{3.9}$$

The maximization of (3.9) for GLM coefficients λ_k can be found by using a weighted, log-linear Poisson regression with weights $1 - z_{ik}^{(s)}$ (McCullagh and Nelder, 1989), yielding $\lambda_k^{(s+1)}$. While the parameter for (3.9) can be maximized over a gradient yielding $\psi_k^{(s+1)}$ (Lambert, 1992).

3.4 Comparing zero-inflated Models

Until recently the usual test for comparing zero-inflated to non-zero inflated models has been the Vuong Test for non-nested models (see Vuong (1989)). However recent work has shown the misuse of this test for zero inflation (see Wilson (2015)). Wilson and Einbeck (2018) show that it is sufficient to test for zero-modificiation in the form of a liklihood ratio test. The test is defined as follows

$$H_0: \quad \psi_k = 0 \qquad vs. \qquad H_a: \quad \psi_k \neq 0$$

$$\varphi = -2\Big[l(\tilde{\lambda_k}; y, \boldsymbol{x}) - l(\lambda_k, \psi_k; y, \boldsymbol{x})\Big] \tag{3.10}$$

where $l(\tilde{\lambda_k}; y, \boldsymbol{x})$ is the log-liklihood of a single component GCWM Poisson model on Ω_{Z_k} parameterized by $\tilde{\lambda_k}$. Recall that ψ_k is the zero-inflation paramater of the kth parition. The test statistic φ is shown to be distributed Chi-square with m degrees of freedom (χ^2_m) and $\alpha = 0.10$ (Wilson and Einbeck, 2018).

4 Application

4.1 Data

We illustrate the proposed methodology on the French motor severity and frequency datasets by policy. These datasets are available as part of the **R** package `CASdatasets` developed by (**?**) and previously used in the book *Computational Actuarial Science with R* by Charpentier (2014). The book demonstrated various GLM modeling approaches for fitting frequency and severity of this data. Miljkovic and Fernndez (2018) used the same datasets to review two mixture-based clustering approaches for modeling unobserved heterogeneity in Region 24 of this insurance portfolio. **?** introduced a Bayesian approach in ratemaking for a new territory considering also policies in French Region 24. French automobile portfolio consists of 413,169 motor third-party liability policies with the associated risk characteristics. The loss amounts by policy ID are also provided.

Table 1: The description of variables in the French Motor Third-Part Liability dataset.

Attribute	Description
Policy ID	Unique identifier of the policy holder
Claim Nb	Number of claims during exposure period (0,1,2,3,4)
Exposure	The exposure of policy in years (0–1.5)
Power	Power level of car ordered categorical (12 levels)
Car Age	Car age in years
Driver Age	Age of a legal driver
Brand	Car brands (7 types)
Gas	Diesel or Regular
Region	Regions in France (10 classifications)
Density	Number of inhabitants per km^2
Loss Amount	Portion of claim the insurance policy pays

4.2 Analysis and Results

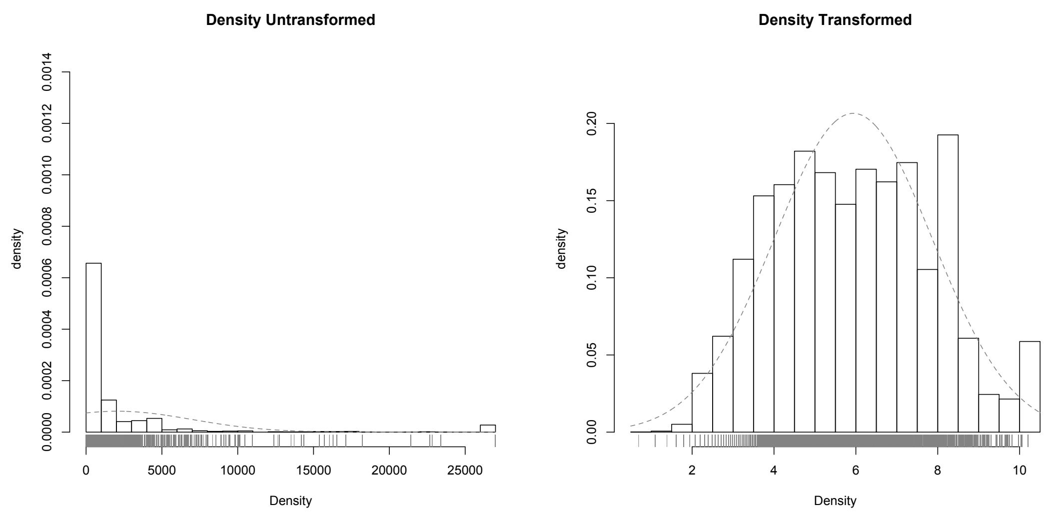
4.2.1 Modelling Severity

In this section, we show the results from modeling French motor losses. We consider the following covariates: density, driver age, car age, power, gas, and region. The model that was fitted is defined with the following equation where $\epsilon \sim \mathcal{N}(0, \sigma)$, and

$$LossAmount = Density + CarAge + DriverAge + Region + Power + Gas + \epsilon. \tag{4.1}$$

The canonnical log-link is used for the GLM (4.1). Similarly to Miljkovic and Fernndez (2018), car age is modelled as a categorical variable with five categories: $[0, 1)$, $[1, 5)$, $[5, 10)$, $[10, 15)$,and $15+$. Additionally, driver age is modelled as a categorical variable with five categories: $[18, 23)$, $[23, 27)$, $[27, 43)$, $[43, 75)$, and $75+$. Power is modelled into three categories as in Charpentier (2014): DEF, GH, and other.

Figure 1: Density variable: Left figure shows the fit when Gaussian distribution is imposed (CMW approach) to highly skewed data. Right figure shows the fit when log-normal assumption is applied (GCWM approach).



Beginning with the continuous covariate $Density$, we want to inspect the shape of its univariate data to see if it follows Gaussian distribution. The left side of Figure 2 clearly revels that the the $Density$ is rather skewed right with several observations that report high value of density. This indicates a need for a transformation. With the log-normal assumption, the $Density$ is transformed which improves the fit (see the right side of Figure 2) on the data.

The result of the transformation is a better AIC and BIC. Table 2 shows a considerable difference in BIC and AIC comparing CWM and GCWM. The five component CWM with a BIC of 285,268 is significantly higher than the four component GCWM with a considerably lower BIC of 89,568.

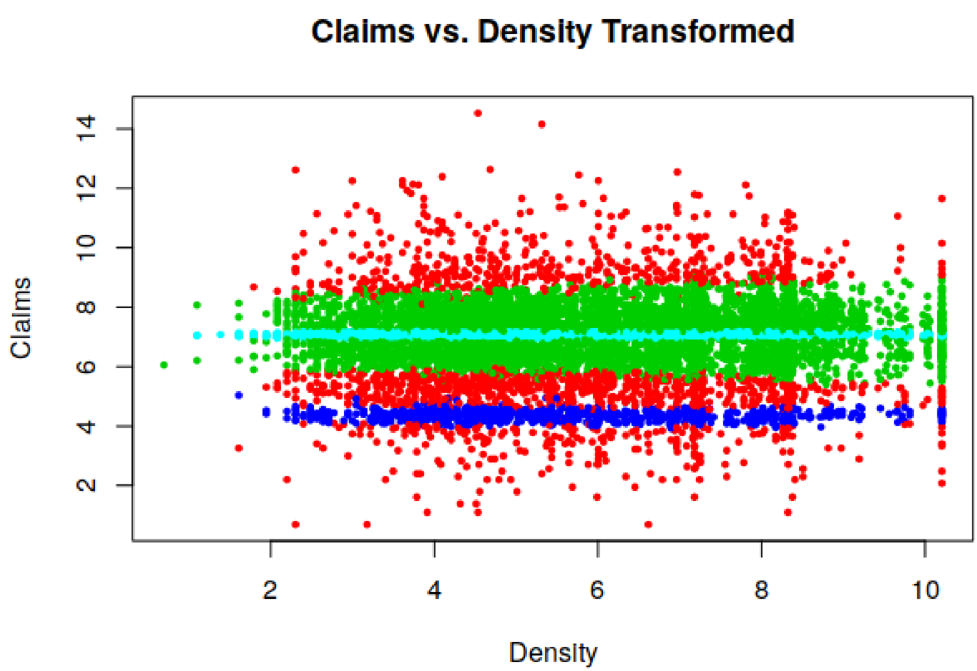
4.2.2 Driver Analysis of Severity Model

We now investigate the results of GCWM in relation to the valuation of risk. For practical uses, finding clusters allows us to create different classifications of risk for various fields of drivers. The following GCWM allows to cluster different drivers in groups allowing one to assign different rates to different clusters.

Table 2: Comparing AIC and BIC for CWM verses GCWM models.

Model	k	AIC	BIC
CWM	1	$352,470$	$352,661$
	2	$314,560$	$314,949$
	3	$301,223$	$301,812$
	4	$287,020$	$287,808$
	5	$\mathbf{284,283}$	$\mathbf{285,268}$
GCWM	1	$111,129$	$111,320$
	2	$90,039$	$90,428$
	3	$89,476$	$90,065$
	4	$\mathbf{88,781}$	$\mathbf{89,568}$
	5	$88,731$	$89,717$

Figure 2: Showing clusters in color for Claims vs Density on a log scale.



After fitting the model, we then take a look at the size of each cluster. The GCWM a function has chosen four components as the best model to represent the data. The size of each cluster is displayed in Table 3. Attention is brought to largest quantity of drivers that are grouped into cluster 4. This accounts for 46% of all

Table 3: Size of clusters for the GCWM a model.

1	2	3	4
1,683	5,766	848	7,093
Red	Green	Blue	Teal

drivers and is fairly concentrated in the center of Figure 2. From the results we can create an insurance model with the following characteristics. Cluster 3 drivers have both low variability and low average cost in claims, thus can be insured with a lower rate than other drivers. From a risk management perspective this is the most ideal case as these claims have very low variance and cost. Cluster 4 drivers have also low variability but a higher average cost, thus they would have a rate higher than cluster 3. Cluster 2 drivers have the next highest cost and variability of the clusters, these drivers are colored in green in Figure 2. The final cluster colored in red has the highest cost and variaiblity in claims out of any of the other clusters. From a risk management perspective this cluster would have the highest rate.

Table 4: Summarized volatility information of each cluster for Claims.

Volatility Level - (Cluster)	Minimum	Mean	Maximum	$\sigma(0.05)$
V1 - (3)	51	79	154	**13.19**
V2 - (4)	1039	3109	1324	**52.07**
V3 - (2)	221	1687	8841	**1284.41**
V4 - (1)	2	9717	2036833	**64835.96**

Table 4 shows a breakdown of the types of drivers, ordered by volatility in descending order. Beginning with V1, these drivers tend to have claims between $51 to $154, with a standard deviation of $13.19, and a mean of $79. That means that these drivers rarely exceed costs and tend to have very low volatility. Moving onto V2, these drivers have the second level of volatility. Drivers in this range tend to have claims anywhere between $1039 to $3109, with a standard deviation of $ 52.07, and a mean of $3109. Proceeding to V3, its volatility in claims is greater than the preceding levels. Drivers in this cluster have claims anywhere between $221 to $8841, with a mean of $1687, and a standard deviation of $1284.41. Finally V4 denotes the level of highest volatility. Claims in this level reach the highest recorded claim of $ 2036833, a mean of $9717, and a standard deviation of $64835.96.

Coefficients of clustered results are used to calculate premiums in car insurance. Table 12 shows the coefficients of the fitted model. The significance codes are defined as $P < 0.001$: (***), $0.001 < P < 0.01$: (**), $0.01 < P < 0.05$: (*), $0.05 < P < 0.10$: (.) pertaining to the P value of the specific coefficient. In each cluster significance varies but overall the majority of coefficients are significant.

To summarize, the drivers have been clustered into four categories with distinct characteristics outlined in Table 4. We have seen how using the results from GCWM, one can create an insurance model based on clustering algorithms with various levels of risk represented in each cluster. GCWM found a group that was the clear majority of drivers, in which the volatility of their claims was extremely low regardless of $Density$ or $DriverAge$ This finding shows that GCWM may potentially find unique clusters that are otherwise hidden within the data.

4.2.3 Modeling Claims Frequency

In this section, we model frequency of the French motor claims. We consider the covariates density, driver age, car age, exposure and power. The choice of covariates stems from the previously modelled single component ZIP (Charpentier, 2014). The GCWM is modelled with the linear formula

$$ClaimNb = Density + Exposure + Power \quad | \quad Exposure + CarAge \tag{4.2}$$

where $Density$, $Exposure$, and $Power$ models Poisson, while $Exposure$ and $CarAge$ models Bernoulli. As in Section 4.2.1, we also impose a log-normal assumption on the $Density$ covariate. After fitting the model, GCWM has found two zero-inflated components and one Poisson component as the best model to represent the data. The size of each cluster is displayed in Table 5. We note a fairly even spread of the size across the three clusters.

Table 5: Size of clusters for the GCWM a model.

1	2	3
100492	163503	149174
Red	Green	Blue

Similarly to modelling severity, the GCWM finds clusters with unique characateristics. This is evident when looking at the Claims vs. Density plot in Figure 3. We see that the GCWM has split up the drivers into three groups based on the Density of cities. Table 6 shows that cluster 2 drivers live in the least dense cities with a mean of 7.86 km on the log scale. Followed by clusters 3 and 1 with a mean of 5.23 km and 3.38 km respectively.

Table 13 shows a summary of the coefficients for the zero-inflated model. The significance codes are the same as of Table 12. In each cluster we can see that the majority of the coefficients are signficant particularly in the zero-count model for Bernoulli. Cluster 1 was selected to be strictly a Poisson model by the liklihood ratio test defined in (3.10). In summary we see that the GCWM can account for zero-inflated data in pricing.

Figure 3: Showing clusters in color for Frequency vs Density under lognormal assumptions.

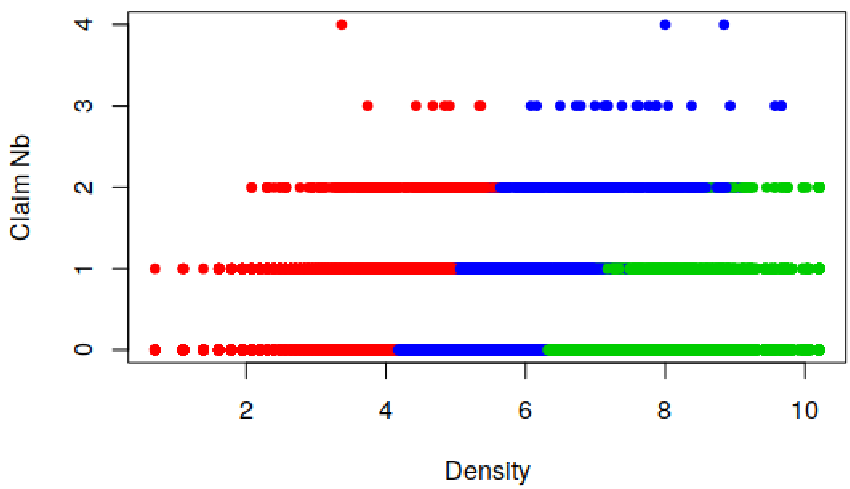


Table 6: Summary of each cluster with log-normal assumptions for the Density covariate.

Cluster - Color	Minimum (km)	Mean (km)	Maximum (km)	σ (km)
1 - Red	0.69	3.38	5.60	0.60
2 - Green	6.29	7.86	10.20	1.03
3 - Blue	4.13	5.23	9.66	0.65

5 Simulation Study

Two simulation studies are conducted to determine the validity of the log-normal assumption and the effectiveness of the Bernoulli-Poisson partitioning method. The first section outlines the need for a non-gaussian assumption for the covariates. The second section shows the classification accuracy and other relevant analysis for the Bernoulli-Poisson method.

5.1 Simulation Study - GCWM

In this section, we show how the proposed methodology works for different simulation settings. The simulation study was generated based on the regression coefficients of the **CASdataset** used in the previous section. The aim of the simulation study was to test the accuracy and ability of both GCWM a and CWM to return estimates of true parameters when one or more of the covariates is lognormal and the other two are Gaussian. This was designed to test both functions in the event when one of the covariates is non-Gaussian. The motivation behind this is fact is that many covariates used in insurance are likely to come from non-Gaussian distribution. Thus this was aimed to test the relevancy of CWM, which treats all covariates as Gaussian.

We define Model 1 as the base line model in which the coefficients were generated for **CASdataset** and reported in upper portion of Table 2. These coefficients were then rounded and treated as true parameters. A simulation with three GLM mixture components was then generated around these true parameters in which the third covariate X_3 was lognormal. Stemming from this, both CWM and GCWM a were run. The GCWM a treats X_3 as a lognormal covariate.

The results for Model 1 were summarized in upper portion of Table 3 based on the performance of the GCWM a approach. The simulation was run 1000 times. We reported the percentage of runs for each predictor and the corresponding intercept in each mixture component under the assumption of 5% error. For example, predictor X_2 in the component 2 of Model 1 reported 90.10% accuracy. This means that 90.1% of the time the true parameter was estimated within 5% error. In this setting, predictor X_1 in the second component was insignificant in the real data set. The purpose of including this parameter in Model 1 was to test the sensitivity of GCWM a for insignificant predictors. In this case, the result of zero is underlined and it means that it has no influence on the response variable in this simulation. Further, we created Models 2, 3, 4 and 5 by altering the parameters of Model 1 by $+30\%$, -30%, $+50\%$, and -50% accordingly and keeping the second covariate of the second component as an insignificant predictor form the **CASdataset** model. This was done to test the accuracy of GCWM a to the sensitivity of coefficients. Based on the results in Table 3, we can see that GCWM a performs well for all simulation settings.

Table 4 provides the summary of the results when CWM was used in the analysis of the same models considered in Table 3. It is not surprising to see that barely any of the simulation runs estimated correctly all parameters as most of the results are zero. This means that the performance of CWM approach is poor in presence of one non-Gaussian covariate which in this case is a log-normal covariate. Similarly to Table 3, Table-4 shows the underlined results pointing to insignificant predictors.

Table 5 provides the summary of Mean Squared Errors (MSE) of each parameter of the models in Table 3 estimated via 1000 simulation runs. The MSE is computed using the following formula MSE(β_i) $=$ $\frac{\sum_i^n (\beta_i - \hat{\beta}_i)^2}{n}$. The MSEs related to the predictor variables for all models and their corresponding components are about zero indicating that GCWM a approach performs well. This is also a result of having a small size coefficients.

Table 7: GCWM a vs CWM Accuracy: Covariate X_3 is treated as log-normal, the rest are Gaussian covariates.

Model	Component	Intercept	X_1	X_2	X_3	Intercept	X_1	X_2	X_3
1	1	93.00%	90.10%	93.00%	93.10%	0.00%	0.00%	0.00%	0.00%
	2	90.10%	0.00%	90.10%	90.10%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.10%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
2	1	89.80%	89.20%	89.80%	89.80%	0.00%	0.00%	4.60%	0.00%
	2	89.20%	0.00%	89.20%	89.20%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	1.70%	0.00%
3	1	100.00%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	2	100.00%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
4	1	88.60%	86.80%	88.60%	87.00%	0.00%	0.00%	0.00%	0.00%
	2	86.90%	0.00%	86.90%	86.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
5	1	85.90%	84.90%	85.60%	85.90%	0.00%	0.00%	0.00%	0.00%
	2	85.00%	0.00%	84.90%	84.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	10.90%	0.00%

Table 8: GCWM a results: the summary of MSE for all parameters used in five models. The covariate X_3 is treated as log-normal a nd the rest are Gaussian. These results correspond to those in Table 3.

Model	Component	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	(11.353)	0.03	(0.00)	3.5	(0.00)	-380	(0.09)
	2	1600	(0.000)	-0.01	(0.00)	1.5	(0.00)	-250	(0.00)
	3	40000	(0.035)	-6.00	(0.00)	-305	(0.00)	1100	(0.47)
2	1	1350	(0.167)	0.04	(0.00)	4.5	(0.00)	-500	(0.03)
	2	2080	(0.001)	0.04	(0.00)	2.0	(0.00)	-325	(0.00)
	3	52000	(0.012)	-8.00	(0.00)	450	(0.00)	14300	(0.01)
3	1	720	(0.001)	0.02	(0.00)	2.5	(0.00)	-266	(0.00)
	2	1100	(0.008)	0.00	(0.00)	1.1	(0.00)	-17511	(0.00)
	3	28000	(0.002)	-4.20	(0.00)	245	(0.00)	7700.	(0.00)
4	1	1650	(13.056)	0.05	(0.00)	5.3	(0.00)	-570	(0.00)
	2	2400	(0.000)	-0.01	(0.00)	2.3	(0.00)	-375	(0.00)
	3	60000	(0.051)	-9.00	(0.00)	-457	(0.00)	16500	(0.00)
5	1	500	(1.115)	0.02	(0.00)	2.0	(0.00)	-190	(0.05)
	2	800	(0.003)	0.00	(0.00)	0.8	(0.00)	-120	(0.00)
	3	20000	(0.000)	-3.00	(0.00)	-150	(0.00)	5500	(0.00)

Table 6 provides the summary of Mean Squared Errors (MSE) of each parameter of the models in Table 4 estimated via 1000 simulation runs. In contrary to the results reported in Table 5, these results in Table 6 are significantly different. We can observe that the MSEs for most of the Models and their corresponding

Table 9: CWM results: the summary of MSE for all parameters used in five models. All three covariates are treated as Gaussian. These results correspond to those in Table 4.

Model	Component	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	$(\cdot)$	0.03	$(\cdot)$	3.5	$(\cdot)$	-380	$(\cdot)$
	2	1600	$(\cdot)$	-0.01	$(\cdot)$	1.5	$(\cdot)$	-250	$(\cdot)$
	3	40000	$(\cdot)$	-6.00	$(\cdot)$	-305	$(\cdot)$	1100	$(\cdot)$
2	1	1350	$(\cdot)$	0.04	$(\cdot)$	4.5	$(\cdot)$	-500	$(\cdot)$
	2	2080	$(\cdot)$	0.04	$(\cdot)$	2.0	$(\cdot)$	-325	$(\cdot)$
	3	52000	$(\cdot)$	-8.00	(0.006)	450	(44.1)	14300	$(\cdot)$
3	1	720	$(\cdot)$	0.02	$(\cdot)$	2.5	$(\cdot)$	-266	$(\cdot)$
	2	1100	(65.814)	0.00	$(\cdot)$	1.1	$(\cdot)$	-17511	$(\cdot)$
	3	28000	$(\cdot)$	-4.20	$(\cdot)$	245	$(\cdot)$	7700.	$(\cdot)$
4	1	1650	$(\cdot)$	0.05	$(\cdot)$	5.3	$(\cdot)$	-570	$(\cdot)$
	2	2400	$(\cdot)$	-0.01	$(\cdot)$	2.3	$(\cdot)$	-375	$(\cdot)$
	3	60000	$(\cdot)$	-9.00	$(\cdot)$	-457	$(\cdot)$	16500	$(\cdot)$
5	1	500	$(\cdot)$	0.02	$(\cdot)$	2.0	$(\cdot)$	-190	$(\cdot)$
	2	800	$(\cdot)$	0.00	$(\cdot)$	0.8	$(\cdot)$	-120	$(\cdot)$
	3	20000	$(\cdot)$	-3.00	(0.003)	-150	(4.7)	5500	$(\cdot)$

components are not generated at all and as such they are shown as $(\cdot)$. This is not surprising because Table 4 shows the accuracy of CWM is not good when attempting to model non-Gaussian predictors as Gaussian.

In summary, our simulation results showed good performance of GCWM a approach in modeling non-Gaussian covariates. More specifically, these results show high accuracy when covariates are log-normal. In contrary, CWM fails to estimate parameters accurately when the Gaussian assumption is violated.

5.2 Simulation Study - Bernoulli-Poisson Partitioning

In this section we show how the Bernoulli-Poisson partitioning (BP) method behaves under different conditions. The components were genereated under similar coefficients taken from the **CASDatasets** package. The coefficients were rounded and treated as true parameters to which data was generated from. The mean and standard deviation of the covariates within each component was also taken into account when generating data. The first simulation examines the performance of the GCWM a model for classification. We generate three components each with sample size $N = 1000$ for a total of 3000 simulated points. The model generated is similar to the mean and standard deviations of Table 6. Consider three simulated covariates and the following GLM model

$$\text{SimClaims} \sim \text{SimDriverAge} + \text{SimLogDensity} + \text{SimCarAge}.$$

Here the GCWM a model is fitted to the simulated data and used to classify into three components. The misclassification rate is calculated by the proportion of true labels placed in other components by the GCWM a model. The results of the simulation is based on the generated dataset are presented in Table **??**. The total misclassification rate is 1.8% and the majority of misclassified components are between components two and three.

Table 10: Misclassfication rate and label comparison of generated data.

True Labels	Classified			Misclassification Rate (%)
	1	2	3	
1	992	3	5	0.80
2	0	990	10	1.00
3	15	20	965	3.50
Overall Misclassification Rate				1.80 %
Average Purity				98.23 %
Adjusted Rand Index				0.9479

$$n_{ij} = \text{ across diagonal }, \qquad a_i = \text{ row sums }, \qquad b_j = \text{ column sums}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \qquad AP = \frac{1}{N}\sum_i n_{ij}$$

The experiment is expanded further to show how Bernoulli-Poisson partitioning behaves over 1000 runs and under two different conditions. The first condition is defined as follows. The mean and standard deviations are taken as given by the estimated ZIP components from the **CASDataset**. The second condition involves adjusting the means of two of the covariates so they are closer to each other. The goal is to show that the BP-method holds its use even when means among covariates are close. Conditions are divided into two categories. N is considered normal, where the covariate means are taken directly from the sample data. C is considered to be “close”, where the covariate means are manipulated so that they are closer to each other within some degree. This is a common problem in classification where if the means among two different components are close, then misclassification rate increases [Lim et al. (2014)]. Experiment 2 defines the use of 3 different partitioning methods to initialize a zero-inflated model. Poisson method assumes that the presence of non-zeros will provide a better partitioning of the data-set. Bernoulli assumes that the presence of excess zeros will determine the best partitioning of the data-set . Finally the BP-Method assumes that both methods are weighed equally and therefore both must be taken into account when partitioning the dataset. The mean and standard deviation of each measurement is provided in Table 11.

Table 11: Experiment 2: mean and standard deviations for each statistic comparing each method.

Type	Condition	Poisson (σ)	Bernoulli (σ)	BP-Method (σ)
Misclassification Rate	N	1.70% (6.00)	1.60% (6.00)	1.10% (0.02)
	C	5.00% (7.00)	6.00% (2.00)	7.00% (4.00)
Average Purity	N	98.87% (2.00)	98.91% (2.25)	99.18% (0.81)
	C	95.38% (4.00)	94.55% (1.00)	96.95% (0.48)
Adjusted Rand Index	N	0.9662 (0.07)	0.9677 (0.07)	0.9729 (0.0217)
	C	0.8706 (0.08)	0.8366 (0.04)	0.8538 (0.0453)

Several findings are concluded from Table 11. Under condition N, the BP method shows better performance in error and is found to be less sensitive than other methods with an error rate of 1.10% and a standard deviation of 0.02%. Further findings show that when condition C is imposed then Bernoulli has better performance in terms of accuracy. The ARI shows good measurements overall however the BP-Method under condition N has a very good ARI with a small standard deviation. The Average Purity of the BP-Method is the best out of all other methods, which is relevant to estimating coefficients accurately for optimization.

6 Conclusion

In this paper, we extend the class of generalized linear mixture CWM models by accomplishing two main goals. First, we proposed the methodology that allows for continuous covariates to follow a non-Gaussian distribution. Imposing Gaussian distribution on a skewed data may result in an suboptimal model fit. Second, we proposed a new Poisson CWM methodology that uses Bernoulli-Poisson partitioning and allows for implementation of zero-inflated Poisson CWM model (ZI-GCWM). We call our proposed model class GCWM, which reflects two extensions made to the existing CWM class of models.

Our proposed GCWM models allow for great applications in predictive modeling of insurance claims by overcoming a few limitations of the current CWM models. The ZI-GCWM allows for finding clusters within claims frequency which is an important information in risk classification and modeling of claims frequency. Further, some insurance rating variables used in the predictive modeling of severity claims may not strictly follow Gaussian assumptions, e.g. driver's age or car age (treated as continuous covariates). An adequate extension to non-gaussian covariates can be considered to relax current assumptions and improve the model fit. We demonstrated that there is a need for a log-normal assumption which can be considered easily to improve the model fit.

The results of our extensive simulation study showed the excellent performance of the proposed models in case of modeling non-Gaussian covariates. We found that current CWM model fails to estimate the parameters accurately when the Gaussian assumption is violated. The GCWM a shows significant improvement in the model fit over the CWM model based on AIC and BIC criteria. We also tested Bernoulli-Poisson partitioning of zero-inflated GCWM under different conditions and found that our proposed partitioning method has a very low misclassification rate, high average purity, and high average rand index.

Our approach is relevant to the actuarial pricing and risk management when current practices are based on implementation of various GLM models. Further extension of this work may incorporate modifications of the CWM family to allow for modeling limited depended variable or the right-censored data structure (refer to **?** and **?**).

7 Appendix

7.1 Derivation of the Log-normal Distribution

Consider a random variable U having univariate log-normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. Have $u \in \mathbb{R}_+$, then the probability density function of random variable U is defined as [1]

$$\mathcal{LN}(u; \mu, \sigma) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln u - \mu)^2}{2\sigma^2}\right].$$

Further, if random variable X is normally distributed i.e. $X \sim \mathcal{N}(x; \mu, \sigma)$, then $U := \exp(X) \sim \mathcal{LN}(u; \mu, \sigma)$. To see this, let $p_U(u)$, and $p_X(x)$ be the probability density functions of U and X respectively. By the change of variables theorem (see Murphy and Bach (2012) section 2.6.2.1) the density $p_U(u)$ is derived as

$$p_U(u) = p_X(\ln u)\frac{\partial}{\partial u}\ln u = p_X(\ln u)\frac{1}{u} = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln u - \mu)^2}{2\sigma^2}\right].$$

We extend to a log-normal multivariate case where the random variable $\boldsymbol{U}$ is parameterized by $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in {\mathbb{R}_+}^p$.

Lemma 7.1. *Let the random variable $\boldsymbol{X}$ have multivariate normal distribution ie. $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{U} := \exp(\boldsymbol{X}) \sim f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here have $\boldsymbol{u} \in {\mathbb{R}_+}^p$ and the probability density function f^U is*

$$f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\prod_{i=1}^N u_i)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu})^{'}\boldsymbol{\Sigma^{-1}}(\ln \boldsymbol{u} - \boldsymbol{\mu})\right].$$

Proof. Let $f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f^X(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the probability density functions of $\boldsymbol{U}$ and $\boldsymbol{X}$ respectively. By the multivariate change of variables theorem (see Murphy and Bach (2012) section 2.6.2.1), we derive the log-normal distribution, where $|\det J_{\ln}(u)|$ is the absolute value of the determinant for the Jacobian of the multivariate transformation $\ln(\boldsymbol{U}) = \boldsymbol{X}$. Hence,

$$\begin{aligned}
|\det J_{\ln}(\boldsymbol{u})| &= \prod_{i=1}^{n} u_i^{-1}, \text{ and} \\
f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= f^X(\ln \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})|\det J_{\ln}(u)| \\
&= f^X(\ln \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\prod_{i=1}^{n} u_i^{-1} \\
&= \frac{1}{(\prod_{i=1}^N u_i)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu})^{'}\boldsymbol{\Sigma^{-1}}(\ln \boldsymbol{u} - \boldsymbol{\mu})\right].
\end{aligned}$$

□

[1] For full definition see Johnson et al. (1995)

Table 12: Summary of coefficients for severity clusters.

	V1	(Red)		V2	(Green)		V3	(Blue)		V4	(Teal)	
Coef	Estimate	Error	P	Estimate	Error	P	Estimate	Error	P	Estimate	Error	P
Intercept	7.876	0.137	***	7.180	0.061	***	4.673	0.014	***	7.077	0.003	***
Density	-0.031	0.009	***	0.005	0.004		-0.011	0.001	***	0.002	0.002	
C2	-0.172	0.080	*	0.064	0.034	.	0.020	0.001	**	0.008	0.002	***
C3	-0.396	0.080	***	0.108	0.034	**	0.010	0.007		0.003	0.002	.
C4	-0.642	0.081	***	-0.033	0.035		0.034	0.007	***	0.005	0.002	**
C5	-0.500	0.090	***	0.066	0.039	.	0.069	0.007	***	0.011	0.002	***
D2	-0.535	0.083	***	-0.168	0.038	***	-0.217	0.009	***	-0.006	0.001	***
D3	-0.607	0.084	***	-0.241	0.038	***	-0.205	0.009	***	-0.008	0.001	***
D4	-0.390	0.099	***	-0.122	0.045	**	-0.200	0.0106	***	-0.009	0.002	***
D5	0.123	0.101		0.035	0.047		-0.138	0.010	***	-0.002	0.002	
R23	0.003	0.131		-0.016	0.053		-0.001	0.012		0.002	0.006	
R24	-0.232	0.054	***	-0.017	0.023		-0.102	0.005	***	-0.015	0.013	***
R25	0.144	0.096		-0.184	0.043	***	-0.065	0.009	***	-0.016	0.024	***
R31	-0.009	0.073		0.055	0.031	.	-0.141	0.008	***	-0.003	0.018	.
R52	-0.303	0.064	***	0.012	0.028		-0.142	0.006	***	-0.015	0.038	***
F53	-0.153	0.063	*	0.095	0.028	***	-0.012	0.006	*	-0.014	0.001	***
R54	-0.222	0.082	**	0.074	0.037	*	-0.122	0.007	***	-0.015	0.002	***
R72	-0.098	0.072		0.175	0.031	***	-0.081	0.007	***	-0.007	0.002	***
R74	-0.236	0.142	.	-0.114	0.067	.	0.466	0.016	***	-0.019	0.003	***
P-FGH	0.123	0.033	***	0.012	0.015		0.001	0.003		0.002	0.001	*
P-Other	0.131	0.045	**	0.075	0.020	***	0.012	0.003	**	0.005	0.001	***
GR	-0.095	0.031	**	-0.029	0.014	*	0.005	0.002	.	-0.005	0.001	***

Table 13: Summary of coefficients for frequency clusters.

	Cluster 1	(Red)		Cluster 2	(Green)		Cluster 3	(Blue)	
Coef	Estimate	Error	P	Estimate	Error	P	Estimate	Error	P
(Intercept)	-10.199	0.109	***	-7.217	0.216	***	-13.694	0.099	***
Density	1.860	0.027	***	0.442	0.013	***	1.771	0.013	***
Exposure	0.825	0.040	***	0.726	0.192	***	0.745	0.046	***
P-GH	-0.063	0.030	*	-0.004	0.035		0.013	0.030	
P-Other	-0.019	0.042		0.049	0.043		0.107	0.040	**
(Intercept)				2.424	0.188	***	1.886	0.356	***
Exposure				-5.871	0.758	***	16.520	2.990	***
C2				-0.541	0.176	**	-0.922	0.368	*
C3				-1.217	0.201	***	-2.263	0.553	***
C4				-1.265	0.225	***	10.805	72.677	
C5				-0.808	0.266	**	-6.355	30.867	

References

Bermúdez, L., Karlis, D., 2012. A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. Computational Statistics and Data Analysis 56 (12), 3988–3999.

Biernacki, C., Celeux, G., Govaert, G., Jul 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (7), 719–725.

Brown, G., Buckley, W., 2015. Experience rating with poisson mixtures. Annals of Actuarial Science 9 (02), 304–321.

Charpentier, A., 2014. Computational Actuarial Science with R. CRC press.

Czado, C., Kastenmeier, R., Brechmann, E., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. Scandinavian Actuarial Journal 4, 278–305.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39, 1–38.

Frees, E.W., D. R., Meyers, G. e., 2014. Predictive modeling applications in actuarial science. Cambridge University Press 1.

Frees, E., Lee, G., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 4 (1), 4.

Frees, E., Wang, P., 2006. Copula credibility for aggregate loss models. Insurance: Mathematics and Economics 38 (2), 360–373.

Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics 70, 205–215.

Gershenfeld, N., 1997. Nonlinear inference and cluster-weighted modeling. Annals of the New York Academy of Sciences 808 (1), 18–24.

Gershenfeld, N., 1999. The nature of mathematical modeling. Cambridge university press.

Gershenfeld, N., S. B., Metois, E., ., 1999. Cluster-weighted modelling for time-series analysis. Nature 397 (67171), 329–332.

Ibrahim, J., Laud, P., 1991. On bayesian analysis of generalized linear models using jeffreys's prior. Journal of the American Statistical Association 86 (416), 981–986.

Ingrassia, S., M. S., Punzo, A., 2014. Model-based clustering via linear cluster-weighted models. Computational Statistics & Data Analysis 71, 159–182.

Ingrassia, S., M. S., Vittadini, G., 2014. Local statistical modeling via a cluster-weighted approach with elliptical distributions. Journal of classification 29 (3), 363–401.

Ingrassia, S., P. A. V. G., Minotti, S., 2015. The generalized linear mixed cluster-weighted model. Journal of Classification 32 (1), 85–113.

Johnson, N., Kotz, S., Balakrishnan, N., 1995. Continuous univariate distributions. No. v. 2 in Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons.
URL `https://books.google.ca/books?id=0QzvAAAAMAAJ`

Krämer, N., Brechmann, E., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. Insurance: Mathematics and Economics 53 (3), 829–839.

Lambert, D., 02 1992. Zero-inflated poisson regression, with an application to defects in manufacturing 34, 1–14.

Lee, S. C. K., Lin, X. S., 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. North American Actuarial Journal 14 (1), 107–130.

Lim, H., Li, W., Yu, P., 03 2014. Zero-inflated poisson regression mixture model 71, 151?158.

McCullagh, P., Nelder, J., 1989. Generalized linear models. Vol. 37. CRC press.

McLachlan, S., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, Hobuken, NJ.

McNicholas, P., Murphy, T., McDaid, A., Frost, D., 2010. Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. Computational Statistics Data Analysis 54 (3), 711 – 723, second Special Issue on Statistical Algorithms and Software.
URL `http://www.sciencedirect.com/science/article/pii/S0167947309000632`

Miljkovic, T., Fernndez, D., May 2018. On two mixture-based clustering approaches used in modeling an insurance portfolio. Risks 6 (2), 57.
URL http://dx.doi.org/10.3390/risks6020057

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics.

Murphy, K., Bach, F., 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machi. MIT Press.
URL https://books.google.ca/books?id=NZP6AQAAQBAJ

Punzo, A., Ingrassia, S., ., 2014. Parsimonious generalized linear Gaussian cluster-weighted models. Springer International Publishing.

Shi, P., Feng, X., Ivantsova, A., 2015. Dependent frequencyseverity modeling of insurance claims. Insurance: Mathematics and Economics 64, 417–428.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., Lin, S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. ASTIN Bulletin 45 (3), 729–758.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.
URL http://www.jstor.org/stable/1912557

Wedel, M., 2002. Concominat variables in finite mixture modeling. Statistica Neerlandica 56 (3), 362–375.

Wedel, M., De Sabro, W., 1995. A mixture likelihood approach for generalized linear models. Journal of Classification 12 (3), 21–55.

Wilson, P., 02 2015. The misuse of the vuong test for non-nested models to test for zero-inflation 127.

Wilson, P., Einbeck, J., 2018. A new and intuitive test for zero modification. Statistical Modelling.
URL http://hdl.handle.net/2436/621356